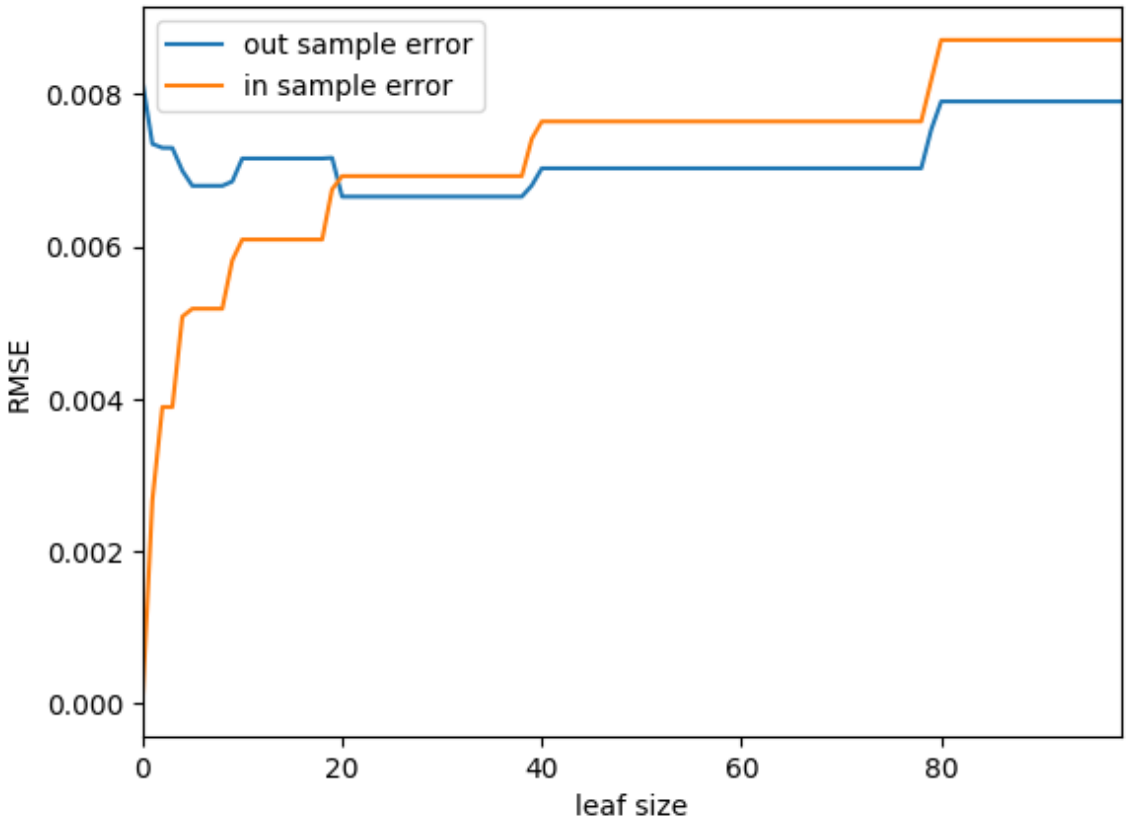
Summer 2019 Project 3: Assess Learners

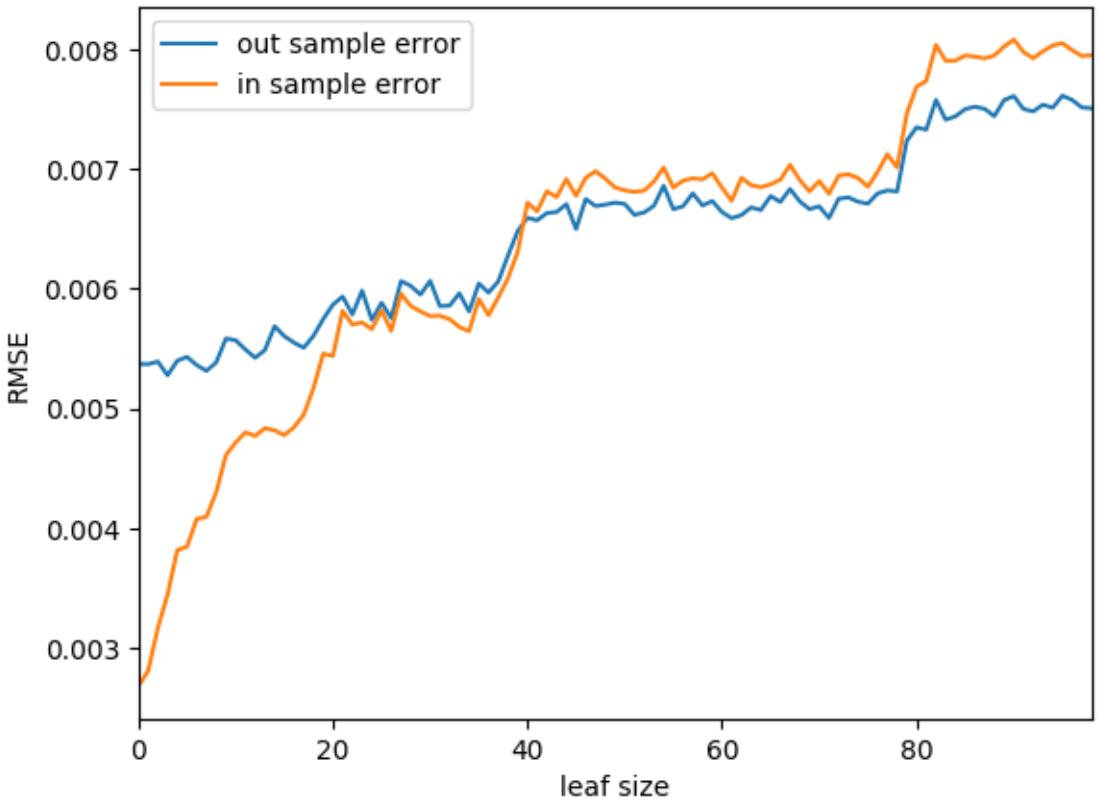
By Nan Mao (nmao7)
GT ID: 903363914

1. Does overfitting occur with respect to leaf_size? Use the dataset istanbul.csv with DTLearner. For which values of leaf_size does overfitting occur? Use RMSE as your metric for assessing overfitting. Support your assertion with graphs/charts. (Don't use bagging).



Overfitting occurs when in sample error is really small but out sample error is very big. The above plot compared the in sample error and out sample error (calculated as RMSE) from Decision Tree when the leaf size in range of 1 to 100. At the left part of the plot, it clearly shows that when leaf size is very small (about less than 20), the out sample error is much higher than the in sample error. As leaf size increases towards to the right part of the plot, the in sample error increases and the out sample error decreases.

2. Can bagging reduce or eliminate overfitting with respect to leaf_size? Again use the dataset istanbul.csv with DTLearner. To investigate this, choose a fixed number of bags to use and vary leaf_size to evaluate. Provide charts to validate your conclusions. Use RMSE as your metric.



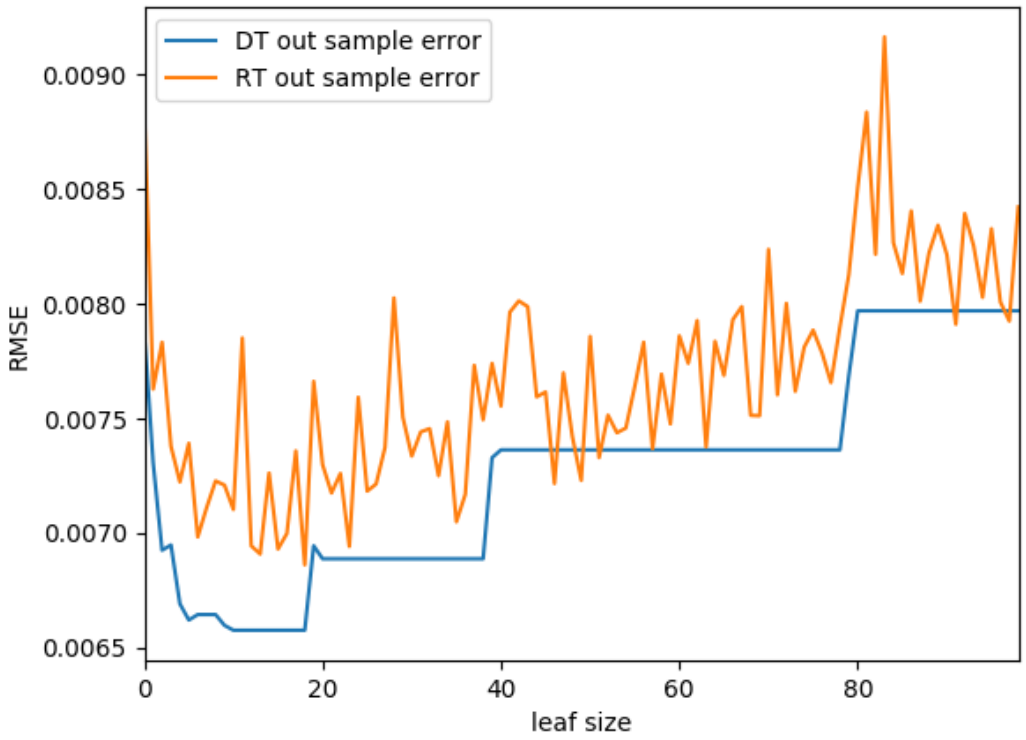
The above plot shows the in sample RMSE and out sample RMSE from BagLearner using a DTLearner and 20 bags. From this plot, the overfitting does still occur at small value of leaf size (about less than 20). While the good thing is bagging reduce the value of RMSE from 0.007 at leaf size of 20 (observed from plot in question 1) to 0.0055 at same leaf size. So bagging algorithm can reduce overfitting by decreasing the error.

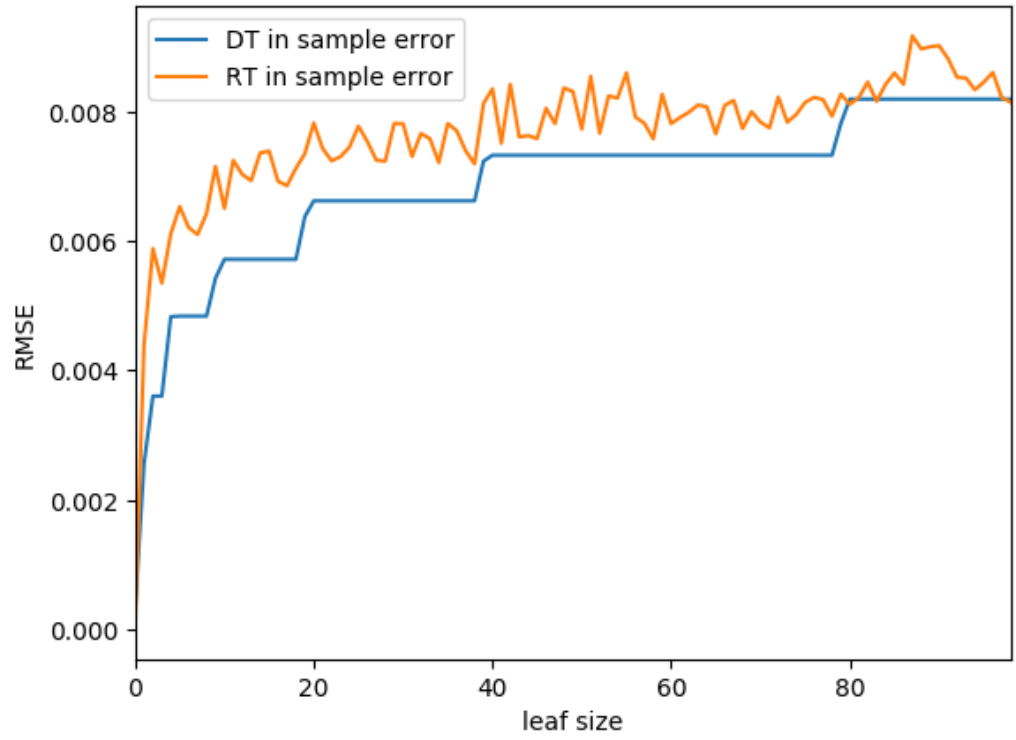
3. Quantitatively compare "classic" decision trees (DTLearner) versus random trees (RTLearner). In which ways is one method better than the other? Provide at least two quantitative measures. Important, using two similar measures that illustrate the same broader metric does not count as two. (For example, do not use two measures for accuracy.) Note for this part of the report you must conduct new experiments, don't use the results of the experiments above for this.

   I choose to calculate out sample RMSE and in sample RMSE as two quantitative measurements to compare the performance of DTLeanrer and RTLearner.

   The plot below shows the comparison between DTLearner and RTLearner based on their out sample RMSEs. Below is the statistics generated from that plot.
   <u>statistics from out sample error of DTLearner:</u>
   mean: 0.00702160483874
   max value: 0.0076829846798
   min value: 0.00632846605527
   <u>statistics from out sample error of RTLearner:</u>
   mean: 0.00756484055374
   max value: 0.00909343347905

min value: 0.00651383517838
DTLearner performs better and the line of plot (blue line) go less fluctuated than RTLearner.





The plot above shows the comparison between DTLearner and RTLearner based on their in-sample RMSEs. Below is the statistics generated from the above plot.

statistics from in-sample error of DTLearner:
mean: 0.00652163169881
max value: 0.0079524120669
min value: 0.0
statistics from in-sample error of RTLearner:
mean: 0.00708466643029
max value: 0.00856704700785
min value: 0.0
DTLearner performs better and the line of plot (blue line) go less fluctuated than RTLearner.
In general, I think DTLearner performs better than RTLearner.